NATIONAL FOUNDATION FOR EDUCATIONAL RESEARCH

# Exploring the Importance of Graders in Determining Pupils' Examination Results Using Cross-Classified Multilevel Modelling

**Tom Benton**

Paper presented at the European Conference on Educational Research,
University of Geneva, 14 September 2006

# Exploring the importance of graders in determining pupils’ examination results using cross-classified multilevel modelling

Tom Benton
NFER, The Mere, Upton Park, Slough SL1 2DQ
t.benton@nfer.ac.uk

## Introduction

High stakes testing is a well established part of education around the world. Results of tests are used for a range of purposes including assessment of the extent to which national performance targets have been met, providing information about the performance of individual schools and informing the future teaching of pupils.

With such high stakes being placed on pupil examinations it is desirable to have an objective standard for grading against which all pupils are assessed (see Moss 1994). However, in disciplines such as English where pupils are generally required to write longer answers or essays, graders are required to make subjective judgements about how well a question has been answered. Under such conditions maintaining consistency between graders may be difficult. Different graders may prefer different styles of writing or attach greater weight to different elements of a pupil’s answer. The aim of this paper is to examine the extent to which such variability can occur and to explore the relationship between this variability and the characteristics of pupils.

## The structure of the tests under consideration

This paper looks at grader consistency in the case of a reading test and a writing test for 11 year olds. The reading test consists of a number of short comprehension exercises each made up of a number of short questions. The format of questions in the reading test varies from simple multiple choice items to questions requiring a few sentences in response. Individual questions have between 1 and 3 points available to be awarded.

The writing test consists of a request for two pieces of writing (one shorter and one longer) on particular subjects. Assessment is then made up of judgements of the following:

- Sentence structure, punctuation and text organisation of shorter task (up to four points may be awarded).
- Composition and effect of shorter task (8 points).
- Sentence structure and punctuation of longer task (8 points).
- Text organisation of longer task (8 points).
- Composition and effect of longer task (12 points).
- Handwriting (3 points).

The sum of the first two of these represents the score for the shorter task (out of 12), and the sum of the last four of these represents the score for the longer task (out of 31). The overall writing score (out of 43) is the sum of these two.

In this paper we will analyse total scores for these reading and writing tests as well as looking at the individual questions and elements that make up these marks.

## Sources of variation

The idea behind the current paper is separate the different influences upon the score that a pupil is awarded by a particular grader. We propose that the overall variability in scores (both between graders and between pupils) is comprised of three separate elements described below.

- **Pupil ability** is the most obvious source of variation. More able pupils will tend to get higher marks than less able pupils. If all graders were to consistently agree on what mark to give then this would account for 100% of the variability in pupil scores.
- **Grader leniency** is defined as the extent to which a grader consistently awards higher marks for a question than other graders.
- **Grader variance** is defined as the extent to which graders may be attracted to different elements of pupils responses. In our conceptualisation this is defined as being separate from leniency in that this does not imply that certain graders will consistently award higher or lower marks. It is recognition of the fact that for one pupil a particular grader may advocate a higher number of marks than other graders but may advocate a lower number of marks than other graders for the next pupil.

It is possible to calculate the percentage of variation between scores that is accounted for by each of these. The mathematical formulation that underpins this work is described below.

## Mathematical formulation

The underlying model may be written in the form of a linear equation:

$$Y_{ijk} = m_j + p_{ij} + m_{jk} + e_{ijk}$$

Where:

$Y_{ijk}$ = Score for pupil *i* on item *j* according to grader *k*

$m_j$ = Mean score for item/strand *j*

$p_{ij}$ = Effect of pupil *i* on the score achieved for item/strand *j*

$m_{jk}$ = Effect of leniency of grader *k* on the score achieved for item/strand *j*

$e_{ijk}$ = Effect of grader variance for pupil *i* and grader *k* for item/strand *j*

Models such as the one described above can be fitted using an ANOVA methodology as described in Shavelson and Webb (1991). Alternative methodologies for this decomposition of grader inconsistency are given by Longford (1995) and Bock et al (2002).

In this paper analysis is performed using cross-classified multilevel models (see Hill and Goldstein 1998). In multilevel modelling terminology $m_j$ may be thought of as a fixed effect since we consider one question at a time. $p_{ij}$ and $m_{jk}$ are random effects since the pupils and graders considered in this analysis are simply a sub-sample of all the many potential pupils and graders in the student population as a whole. The modelling procedure estimates that variance of $p_{ij}$, $m_{jk}$ and $e_{ijk}$. From these estimates it is possible to calculate the percentage of variance that is attributable to each source.

Using multilevel modelling for this task has a number of advantages. Firstly the technique does not require a balanced design in order for analysis to work. Provided we have each script assessed by more than one grader the methodology is robust. This is not to say that experimental design should be ignored since clearly some designs are more efficient than others but it is advantageous to have a model that will work in complex scenarios.

A second advantage that will be explored further in this paper is that the model described above can be easily extended to explore the influence of outside variables on grader inconsistency. Suppose for example we were interested in determining is grader inconsistency is greater when grading tests taken by girls. Using multilevel modelling techniques it is now possible to extend our formulation such that:

$$Y_{ijk} \quad = \quad (m_j + b * I_{ij}) \quad + \quad p_{ij} \quad + \quad m_{jk} \quad + \quad e_{ijk}$$

$$\text{Var}(p_{ij}) = s_p^2 + b_p * I_{ij}$$

$$\text{Var}(m_{jk}) = s_m^2 + b_m * I_{ij}$$

$$\text{Var}(e_{ijk}) = s_e^2 + b_e * I_{ij}$$

where $I_{ij}$ is an indicator of whether pupil $i$ is female. An example of this technique will be given later in the paper.

## Methodology

Scripts from 49 pupils were each independently evaluated by nine experienced graders. Graders were not monitored during this process. The object of the analysis was to discover the extent to which different graders award different marks for each item and each pupil.

Analysis was carried out in three stages. Firstly some descriptive statistics were produced to give a broad feel for the extent and severity of inconsistency between graders. The second stage of analysis used cross-classified multilevel modelling to disaggregate grader inconsistency into grader leniency and grader variance. A third stage of analysis extends the multilevel model to explore changes in grader inconsistency across different subgroups of pupils.

**Descriptive statistics**

For each pupil and each item we have considered two measures of disagreement between graders.

- The range, which measures the number of marks between the best mark awarded by any grader and the worst mark awarded by any grader.
- The standard deviation between graders, which is a more inclusive measure of grader dispersion (i.e. it uses all available data not just the best and worst marks).

The results of the descriptive stage of analysis are displayed in tables 1 and 2. For each item within the reading and writing tests as well as for test and sub-test total scores the following statistics are calculated:

- The mean of the range in scores. For example, for the average pupil there is a difference of 4.8 marks between the best score they were awarded for the reading test as a whole and the worst score they were awarded.
- The maximum of the range in scores. This statistic indicates the largest scale of disagreement that occurred for each item. For example there was at least one pupil where there was a disagreement of 11 marks between the best score they were awarded for the reading test as a whole and the worst score they were awarded.
- The mean across pupils of the standard deviations in score across graders. This gives an indication of the extent to which we might expect an average individual's score to change if their test was to be marked by a different grader.

In order to give a sense of scale to these measures it was decided to compare dispersion between graders with dispersion between pupils. It is clearly highly desirable that which pupil takes an exam has a greater on influence on the mark awarded than which grader is grading it. In order to provide this comparison the standard deviation between pupils was calculated for each grader. The mean of these standard deviations is then compared to the mean standard deviation between graders (described above). The ratio between these two numbers was calculated to provide an assessment of the extent to which the influence of pupils on the marks achieved outweighs the influence of graders. The items in the reading test in table 1 are sorted by this ratio. It is clearly desirable that this ratio is a high as possible so that grader inconsistency has as small an effect on test scores as possible.

Table 1 displays the results for each item in the reading test. Questions are grouped into 5 categories:

- Multiple choice
- Tick box which differ from multiple choice in that a number of ticks may be required to gain the marks
- Constrained response where examinees are to perform tasks with a limited number of possible responses such as picking particular words or phrases from some given text

- Short response where examinees are required to a write a sentence to answer the question
- Longer response where several sentences would be required to gain all the marks on offer

Generally speaking grader inconsistency had a small impact on the variation in scores for most reading items with differences between pupils often being over three times higher than differences between graders. Longer response questions tended to have a higher degree of grader inconsistency than constrained response questions or multiple choice items however there were some exceptions. Notably question 14 (where grader inconsistency appeared to have the greatest influence on scores for any one particular item) was an item requiring students to tick boxes. Performing this analysis highlighted the fact that some graders were finding the instructions on which combinations of ticks related to which marks confusing. Having performed this analysis allowed the instructions for graders to be revised to deal with this problem prior to the test being widely distributed.

It is clear from table 2 that grader inconsistency is a much greater problem within the writing test. For example whereas for the reading test the mean standard deviation between pupils is over five times as high as the mean standard deviation between graders for the writing test this ratio is close to 2. Given the more subjective nature of the decisions graders must make to assign marks to pupils this is probably to be expected. Maker inconsistency has the most severe effect on the assessments of pupils' handwriting.

In both tests the impact of grader inconsistency appears to have a smaller effect on total test scores than on individual items within a test. Reasons for this will be discussed further in the next section.

### Cross-classified multilevel modelling

Descriptive analysis has revealed that there is inconsistency between graders. Cross-classified multilevel modelling was used to attempt to disaggregate this into grader leniency and grader variance. In other words we wish to discover whether this inconsistency is caused because some graders tend to be consistently more generous than others or whether this is a more general type of variation.

Analysis was performed using the multilevel modelling package MlWin. Individual marks were grouped firstly according to the pupil that had taken the test and secondly according to the grader who had marked the test. Since all graders marked the results of all pupils this is a non-hierarchical structure and a cross-classified model is required. It is important to note that one of the advantages of performing analysis in this way is that it would be equally effective in less balanced experimental designs.

Results of analysis are shown in table 3. Since multilevel modelling requires marks to have a roughly normal distribution results are not shown for individual items in the reading test but only for total test score. The technique appeared to work relatively well for all elements of scoring for the writing test and so all these are shown.

It can be seen that for the reading test the variation in scores is almost entirely accounted for by the ability of pupils. The remaining variation (which amounts to less than four per cent) is largely the result of grader variance. Statistical tests of whether grader leniency was a significant influence did not provide a significant result. This implies that there is no evidence of graders being consistently biased in the scores they award for the reading test.

The results for the writing test are less encouraging. For each element of the scoring less than 70 per cent of the variation in scores is accounted for by pupil ability. Grader variance accounts for most of the inconsistency between graders and has the relatively highest effect on assessments of handwriting. Less than half of the variation in handwriting scores is attributable to pupil ability. This indicates that graders frequently disagree about the quality of handwriting. The fact that this is classified as grader variance rather than leniency indicates that for different pupils different graders are awarding the highest and lowest marks.

Grader leniency is a lesser influence but has a statistically significant effect for assessments of longer task sentence structure and punctuation, longer task text organisation and for total scores for the longer task and writing test as a whole.

When looking at total scores for the longer task, shorter task and test as a whole it can be seen that the influence of grader variance is greatly reduced. This implies that grader variance cancels out as the scores assigned to various grading elements are totalled. This leads to a relatively greater degree of consistency between graders when we look at overall test scores than when we look at individual elements. Grader leniency does not appear to cancel out as we total the elements of scoring and remains a statistically significant influence. Having said this grader leniency remains the smallest source of variation in scores.

It is also to be expected that the influence of grader variance would cancel out if we were assessing whole classes or schools of pupils rather than individuals. For example, suppose a group of schools were all to take the same test and that a different grader performed the scoring for each school. Although we might expect grader variance to have an impact on the scores of pupils we would not expect this to have much influence on the comparison of the performance of different schools. As a result grader inconsistency should have little influence on the performance of schools as a whole in any testing programme.

## Grader variance within subgroups of pupils

The multilevel model was then extended to explore whether the degree of Grader inconsistency varied within particular subgroups of pupils. Analysis explored the effect of the following pupil characteristics:

- Gender
- Ability in writing as assessed by the pupil's teacher. Teachers assigned levels of 3, 4 or 5 to each pupil.
- Spelling ability of pupil as assessed by separate 20 word spelling test.
- Whether the pupil used extra paper (in addition to the space provided in the answer booklet) in their response to the longer task.

- Whether the pupil wrote in paragraphs to organise their text during the longer task.

The variance of each of leniency, variance and pupil ability were modelled as additive functions of the given background variables. A backwards stepwise procedure was used to remove variables that were not significant at the 5% level. This procedure resulted in the following equations describing the variances.

$$\text{Var(Grader leniency)} = 4.234-0.002244*(\text{Spelling ability})^2$$

$$\text{Var(Grader variance)} = 0.5487+0.88*(\text{Female pupil})+1.291*(\text{Pupil used extra paper})$$

$$+0.8999*(\text{Pupil wrote in paragraphs})+0.01221*(\text{Spelling ability})^2$$

$$\text{Var(Pupil ability)} = 13.22$$

As can be seen there were no significant links between the variance in scores attributed to pupil ability and the given background characteristics. Also the teacher assessment level of each student was not found to have any significant effect on variance.

The equations indicate that grader leniency decreases as the spelling ability of pupils increases. In other words graders show the greatest propensity to be consistently generous or harsh when faced with pupils with poor spelling ability. Grader variance increases with spelling ability indicating the extent of inconsistent variation between graders is at its highest amongst pupils who are good at spelling. These findings are illustrated in figure 1. This shows that there is a strong relationship between the spelling ability of pupils and the amount of variance attributable to various sources. For pupils with low spelling ability almost a quarter of the variation in their scores is attributable to the leniency of the grader. For pupils with high ability in spelling only a sixth of the variation in scores is attributable to the overall leniency of graders whereas almost a quarter is now assigned to grader variance.

As can be seen from the equations above grader variance also increases slightly in cases where the pupil is female, the pupil uses extra paper or the pupil writes in paragraphs. This indicates that although there is no overall increase in how lenient or harsh graders are for these types of pupils there is a greater degree of inconsistency in grading. This is illustrated further in table 4 which shows how the equations above translate into percentages of variances attributable to various sources for different types of pupil. For the types of pupil listed above roughly 20 per cent of the variation in their scores is attributable to grader variance compared to 13 per cent for male pupils who do not use either extra paper or write in paragraphs.

## Conclusions

Whilst it was found that levels of bias between graders tended to be quite low the analysis clearly shows the importance of exploring grader consistency in test design. Even closed-response questions that would be expected to show complete consistency can display higher than expected levels of variation between graders. Using established analysis techniques to capture these problems during the development phase may give test developers the opportunity to deal with some of the causes.

In the case of writing the analysis shows the difficulty in achieving objective measurement of writing ability. Assessments of handwriting displayed particularly large amounts of grader variance.

Cross-classified multilevel models provide an excellent methodology for assessing inconsistency between graders. Not only is it possible to disaggregate inconsistency amongst graders into its constituent parts but also to extend the model to explore how the sources of variation in grading change for different subgroups. This paper has shown how such analysis can be done. This opens up a number of avenues for future research in attempting to identify those particular styles of writing that are most likely to lead to severe inconsistency amongst graders. Understanding how grader inconsistency changes within different sub-populations of pupils and for different styles of writing may help address some of the root causes and bring a greater degree of reliability in the grading of written work.

## References

Bock, R.D., Brennan, R.L. and Muraki E. (2002). ‘The information in multiple ratings’, *Applied Psychological Measurement*, **26**, 364–375.

Hill, P.W. and Goldstein, H. (1998) ‘Multilevel modelling of educational data with cross-classification and missing identification of units’, *Journal of Educational and Behavioral Statistics*, **23**, 117–128.

Longford, N.T. (1995). *Models for Uncertainty in Educational Testing*. New York, NY: Springer Series in Statistics.

Moss, P.A. (1994). ‘Validity in high stakes writing assessment: problems and possibilities’, *Assessing Writing*, **1**, 109–128.

Shavelson, R.J. and Webb, N.M. (1991). *Generalizability Theory: a Primer*. Newbury Park, NJ: Sage.

## Table 1: Descriptive statistics of grader inconsistency for reading test

| Score (Description) | Mean range between graders (across pupils) | Max range between graders (across pupils) | Mean Standard Deviation between graders (across pupils) | Mean SD between pupils (across graders) - Comparison | Pupil:Grader SD Ratio |
|---|---|---|---|---|---|
| Question 14 (tick box) | 1.0 | 2 | 0.4 | 0.6 | 1.4 |
| Question 10 (short response) | 0.7 | 1 | 0.3 | 0.5 | 1.7 |
| Question 13 (longer response) | 1.2 | 2 | 0.5 | 0.8 | 1.8 |
| Question 25 (short response) | 0.8 | 2 | 0.3 | 0.5 | 1.8 |
| Question 6 (short response) | 0.8 | 2 | 0.3 | 0.7 | 2.1 |
| Question 30 (longer response) | 1.2 | 3 | 0.5 | 1.1 | 2.3 |
| Question 16 (short response) | 0.4 | 1 | 0.2 | 0.5 | 2.8 |
| Question 11 (short response) | 0.5 | 2 | 0.2 | 0.6 | 2.8 |
| Question 7 (longer response) | 0.7 | 2 | 0.3 | 0.8 | 2.9 |
| Question 21 (longer response) | 0.3 | 1 | 0.1 | 0.4 | 3.6 |
| Question 20 (constrained response) | 0.5 | 1 | 0.2 | 0.7 | 3.7 |
| Question 4a (short response) | 0.5 | 2 | 0.2 | 0.7 | 3.9 |
| Question 8 (short response) | 0.3 | 1 | 0.1 | 0.5 | 4.0 |
| Question 26 (short response) | 0.5 | 2 | 0.2 | 0.8 | 4.1 |
| Question 3 (short response) | 0.3 | 1 | 0.1 | 0.5 | 4.1 |
| Question 17a (short response) | 0.6 | 2 | 0.2 | 0.9 | 4.4 |
| Question 17b (short response) | 0.3 | 1 | 0.1 | 0.5 | 4.5 |
| Question 9 (short response) | 0.1 | 1 | 0.1 | 0.3 | 6.8 |
| Question 29b (constrained response) | 0.2 | 1 | 0.1 | 0.5 | 8.3 |
| Question 28 (constrained response) | 0.1 | 1 | 0.0 | 0.5 | 11.3 |
| Question 31 (constrained response) | 0.1 | 1 | 0.0 | 0.5 | 12.5 |
| Question 24c (constrained response) | 0.1 | 1 | 0.0 | 0.5 | 15.0 |
| Question 24a (constrained response) | 0.1 | 1 | 0.0 | 0.5 | 15.7 |
| Question 4b (multiple choice) | 0.1 | 1 | 0.0 | 0.4 | 18.5 |
| Question 5 (tick box) | 0.0 | 1 | 0.0 | 0.3 | 28.0 |
| Question 2 (constrained response) | 0.1 | 2 | 0.0 | 0.6 | 32.0 |
| Question 15 (multiple choice) | 0.0 | 1 | 0.0 | 0.4 | 38.0 |
| Question 23 (multiple choice) | 0.0 | 1 | 0.0 | 0.4 | 39.0 |
| Question 22 (multiple choice) | 0.0 | 1 | 0.0 | 0.4 | 40.0 |
| Question 24b (constrained response) | 0.0 | 1 | 0.0 | 0.4 | 44.0 |
| Question 12 (constrained response) | 0.0 | 1 | 0.0 | 0.5 | 45.0 |
| Question 19 (short response) | 0.0 | 1 | 0.0 | 0.5 | 49.0 |
| Question 27 (constrained response) | 0.0 | 1 | 0.0 | 0.5 | 50.0 |
| Question 1 (multiple choice) | 0.0 | 0 | 0.0 | 0.0 | Not applicable |
| Question 29a (constrained response) | 0.0 | 0 | 0.0 | 0.5 | Not applicable |
| Total score for the reading test | 4.8 | 11 | 1.5 | 8.8 | 5.7 |

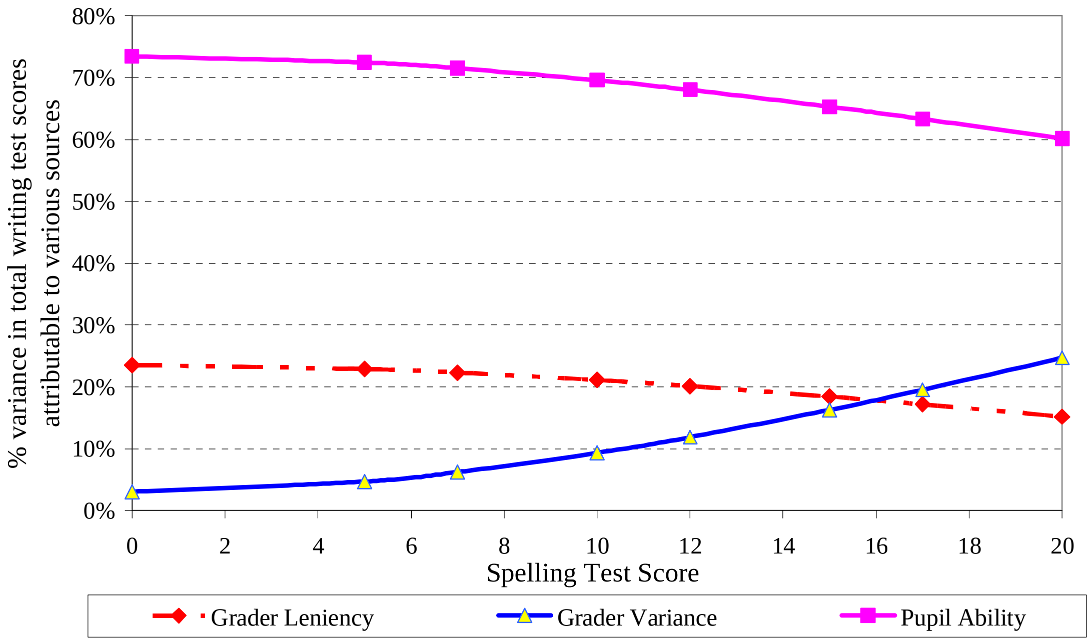
## Table 2: Descriptive statistics of grader inconsistency for writing test

| Score | Mean range between graders (across pupils) | Max range between graders (across pupils) | Mean Standard Deviation between graders (across pupils) | Mean SD between pupils (across graders) - Comparison | Pupil:Grader SD Ratio |
|---|---|---|---|---|---|
| Shorter task sentence structure and punctuation | 1.0 | 2 | 0.4 | 0.8 | 1.9 |
| Shorter task composition and effect | 2.4 | 4 | 0.8 | 1.4 | 1.6 |
| Longer task sentence structure and punctuation | 2.2 | 4 | 0.8 | 1.3 | 1.7 |
| Longer task text organization | 2.2 | 4 | 0.8 | 1.2 | 1.5 |
| Longer task composition and effect | 3.3 | 6 | 1.2 | 2.0 | 1.7 |
| Handwriting | 1.0 | 2 | 0.4 | 0.6 | 1.4 |
| Total score for longer task | 7.2 | 13 | 2.4 | 4.5 | 1.9 |
| Total score for shorter task | 3.3 | 6 | 1.1 | 2.0 | 1.8 |
| Total score for writing test | 9.1 | 18 | 2.9 | 6.1 | 2.1 |

## Table 3: Results of cross-classified multilevel modelling

| Score | Percentage of variance in scores attributable to: | | |
|---|---|---|---|
| | Grader Leniency | Grader Variance | Pupil Ability |
| Total score for reading test | 0.7 | 2.7 | 96.6 |
| Shorter task sentence structure and punctuation | 5.9 | 24.4 | 69.7 |
| Shorter task composition and effect | 9.3 | 24.3 | 66.3 |
| Longer task sentence structure and punctuation | 14.2 | 16.1 | 69.7 |
| Longer task text organization | 14.1 | 24.9 | 61.0 |
| Longer task composition and effect | 8.7 | 24.6 | 66.7 |
| Handwriting | 4.9 | 48.1 | 47.0 |
| Total score for longer task | 10.4 | 16.3 | 73.3 |
| Total score for shorter task | 8.3 | 19.3 | 72.4 |
| Total score for writing test | 9.4 | 12.4 | 78.1 |

## Table 3: The relationship between spelling ability and grader inconsistency



## Table 4: Analysis of grader inconsistency in amongst subgroups of pupils

| Type of pupil | Percentage of variance in total writing test scores attributable to: | | |
|---|---|---|---|
| | Grader Leniency | Grader Variance | Pupil Ability |
| Male pupil of average ability | 19.6 | 13.3 | 67.2 |
| Female pupil | 18.0 | 20.4 | 61.6 |
| Male pupil using extra paper | 17.3 | 23.3 | 59.4 |
| Male pupil using paragraphs | 17.9 | 20.5 | 61.5 |